

E7

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
10 October 2002 (10.10.2002)

PCT

(10) International Publication Number  
**WO 02/079792 A2**

(51) International Patent Classification<sup>7</sup>: **G01S 3/782**,  
H04N 5/232, 7/15, G01S 3/808

Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). GUTTA,  
Srinivas; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

(21) International Application Number: PCT/IB02/00870

(74) Agent: **GRAVENDEEL, Cornelis**; Internationaal  
Octrooibureau B.V., Prof. Holstlaan 6, NL-5656 AA  
Eindhoven (NL).

(22) International Filing Date: 15 March 2002 (15.03.2002)

(25) Filing Language: English

(81) Designated States (*national*): CN, JP.

(26) Publication Language: English

(84) Designated States (*regional*): European patent (AT, BE,  
CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC,  
NL, PT, SE, TR).

(30) Priority Data:  
09/822,121 30 March 2001 (30.03.2001) US

**Published:**

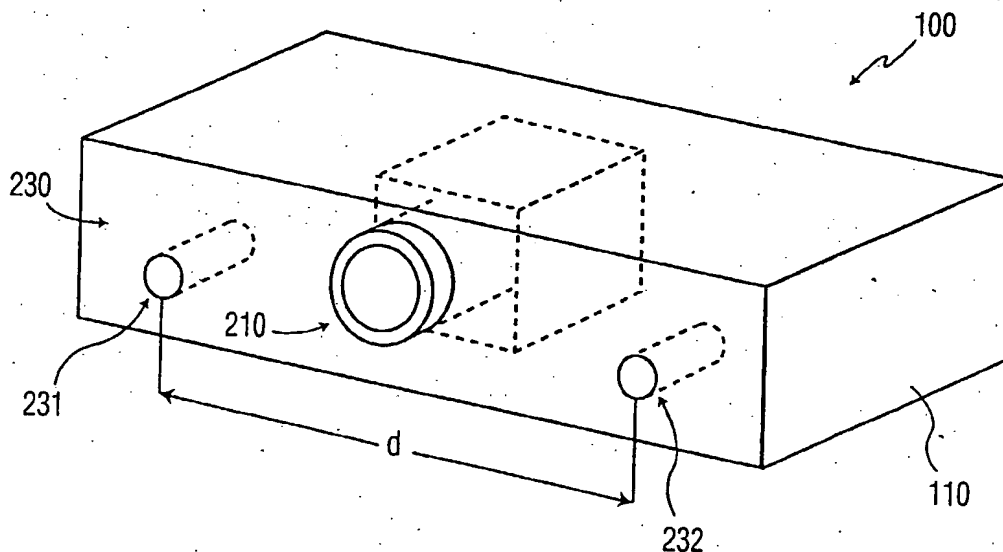
- without international search report and to be republished  
upon receipt of that report
- entirely in electronic form (except for this front page) and  
available upon request from the International Bureau

(71) Applicant: **KONINKLIJKE PHILIPS ELECTRON-  
ICS N.V.** [NL/NL]; Groenewoudseweg 1, NL-5621 BA  
Eindhoven (NL).

(72) Inventors: **COLMENAREZ, Antonio**; Prof. Holstlaan  
6, NL-5656 AA Eindhoven (NL). **STRUBBE, Hugo, J.**;

For two-letter codes and other abbreviations, refer to the "Guid-  
ance Notes on Codes and Abbreviations" appearing at the begin-  
ning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND APPARATUS FOR AUDIO/IMAGE SPEAKER DETECTION AND LOCATOR



(57) Abstract: A method and apparatus for a video conferencing system using an array of two microphones and a stationary camera to automatically locate a speaker and electronically manipulate the video image to produce the effect of a movable pan tilt zoom ("PTZ") camera. Computer vision algorithms are used to detect, locate, and track people in the field of view of a wide-angle, stationary camera. The estimated acoustic delay obtained from a microphone array, consisting of only two horizontally spaced microphones, is used to select the person speaking. This system can also detect any possible ambiguities, in which case, it can respond in a fail-safe way, for example, it can zoom out to include all the speakers located at the same horizontal position.

WO 02/079792 A2

## Method and apparatus for audio/image speaker detection and locator

## BACKGROUND OF THE INVENTION

## 1. Technical Field

The present invention relates to a method and apparatus for a video conferencing system using an array of two microphones and a stationary camera to automatically locate a speaker and electronically manipulate the video image to produce the effect of a movable pan tilt zoom ("PTZ") camera.

## 2. Related Art

Video conferencing systems which determine a direction of an audio source relative to a reference point are known. Video conferencing systems are one variety of visual display systems and commonly include a camera, a number of microphones, and a display. Some video conferencing systems also include the capability to direct the camera toward a speaker and to frame appropriate camera shots. Typically, users of a video conferencing system direct movement of the camera to frame appropriate shots. Existing commercial video conferencing systems use microphone arrays to automatically locate a speaker and drive a pan tilt zoom ("PTZ") video camera. See, for example, (1) Patent Cooperation Treaty Application WO 99/60788, entitled "Locating an Audio Source", and (2) United States Patent No. 5,778,082 entitled "Method and Apparatus for Localization of an Acoustic Source", issued on July 7, 1998 to Chu *et al.*, both documents incorporated herein by reference.

Unfortunately, it is problematic to accurately detect, locate, and track a speaker using an array of only two microphones which function in combination with a stationary video camera. Thus, there is a need for a method and apparatus for a video conferencing system using an array of two microphones to automatically locate a speaker and to then track the speaker using a stationary video camera.

## SUMMARY OF THE INVENTION

Computer vision algorithms are used to detect, locate, and track people in the field of view of a wide-angle, stationary video camera. The estimated acoustic delay obtained from a microphone array, consisting of only two horizontally spaced microphones,

is used to select the person speaking. Assuming that no more than one speaker will be located at exactly the same horizontal position, the acoustic delay between the two microphones provides enough information to unambiguously locate the speaker. The system of the present invention can also detect any possible ambiguities, in which case, it can  
5 respond in a fail-safe way. For example, it can zoom out to include all the speakers located at the same horizontal position.

The audio and video processing steps are performed at an early stage, so that only two microphones and one stationary video camera are needed to locate and track the speaker. This approach reduces the requirements in both hardware and computation, and  
10 improves the overall system performance. For instance, this approach allows the video conferencing system to accurately track moving people regardless of whether they speak or not.

In a first general aspect, the present invention provides a video conferencing system comprising: an image pickup device for generating image signals representative of an  
15 image; an audio pickup device for generating audio signals representative of sound from an audio source; and a multimodal integration architecture system for processing said image signals and said audio signals to determine a direction of the audio source relative to a reference point.

In a second general aspect, the present invention provides a method  
20 comprising the steps of: generating, at an image pickup device, image signals representative of an image; generating, at an audio pickup device, audio signals representative of sound from an audio source; processing the image signals and the audio signals to determine a direction of the audio source relative to a reference point; manipulating the image signals to produce refined image signals; and outputting said refined image signals.

25 In a third general aspect, the present invention provides a video conferencing system comprising: two microphones for generating audio signals representative of sound from a speaker;

a video camera for generating video signals representative of a video image; an electronic pan tilt zoom system for manipulating video images to produce the visual  
30 effects of panning, tilting, and or zooming; a processor for processing the video signals and the audio signals to determine a direction of a speaker relative to a reference point and supplying control signals to the electronic pan tilt zoom system for producing images that include the speaker in the field of view of the camera, the control signals being generated based on the determined direction of the speaker; and

a transmitter for transmitting audio and video signals for video conferencing.

## BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 depicts an exemplary video conferencing system, in accordance with  
5 embodiments of the present invention.

FIG. 2 depicts various functional modules of the video conferencing system of  
FIG. 1, in accordance with embodiments of the present invention.

## DETAILED DESCRIPTION OF THE INVENTION

10 The present invention discloses an apparatus and associated method for a  
video conferencing system using an audio pickup device, such as a microphone array  
consisting of two microphones, and a stationary image pickup device, such as a video  
camera. The video conferencing system of the present invention is able to accurately detect,  
locate, and track a speaker using an array of only two microphones which function in  
15 combination with a stationary video camera.

Referring now to the drawings and starting with FIG. 1, an exemplary video  
conferencing system 100 is shown. Video conferencing system 100 includes a stationary  
video camera 210 and a horizontal array of two microphones 230, which includes a first  
microphone 231 and a second microphone 232, positioned a predetermined distance  $d$  from  
20 one another, and fixed in a predetermined geometry.

Briefly, during operation, video conferencing system 100 receives sound  
waves from a human speaker (not shown) and converts the sound waves into audio signals.  
Video conferencing system 100 also captures video images of the speaker via stationary  
video camera 210. Video conferencing system 100 uses the audio signals and video images  
25 to determine a location of the speaker relative to a reference point, for example, video camera  
210. Based on that direction, video conferencing system 100 can then electronically  
manipulate the video images to effectively pan, tilt, or zoom in or out, the video images from  
stationary video camera 210 to obtain a better image of the speaker.

Generally, the location of the speaker relative to video camera 210 can be  
30 characterized by two values: a direction of the speaker relative to stationary video camera  
210 which may be expressed as a vector, and a distance of the speaker from stationary video  
camera 210. As is readily apparent, the direction of the speaker relative to stationary video  
camera 210 can be used for effectively pointing stationary video camera 210 toward the  
speaker by electronically mimicking a panning or tilting operation of stationary video

camera 210, and the distance of the speaker from stationary video camera 210 can be used for electronically mimicking a zooming operation stationary video camera 210.

It should be noted that in video conferencing system 100 the various components and circuits constituting video conferencing system 100 are housed within an integrated housing 110 in FIG. 1. Integrated housing 110 is designed to be able to house all of the components and circuits of video conferencing system 100. Additionally, integrated housing 110 can be sized to be readily portable by a person. In such an embodiment, the components and circuits can be designed to withstand being transported by a person and also to have "plug and play" capabilities so that the video conferencing system can be installed and used in a new environment quickly.

FIG. 2 schematically shows functional modules of the video conferencing system 100 of FIG. 1. Microphones 231, 232 and stationary video camera 210, respectively, supply audio signals 235 and video signals 215 to a multimodal integrated architecture module 270. Multimodal integrated architecture module 270 includes an audio source localization module 240, a computer vision person detection module 250, and a multimodal speaker detection module 260. An electronic pan tilt zoom (EPTZ) control signal is output from the multimodal speaker detection module 260 and is supplied to an electronic pan tilt zoom system module 220.

A method of operation and associated structure of a typical multimodal integrated architecture module is disclosed in (1) United States Patent Application Serial Number 09/718,255 filed November 22, 2000, entitled "Candidate-level Multimodal Integration Systems"; and (2) United States Patent Application Serial Number 09/548,734 filed April, 13 2000, entitled "Method And Apparatus For Tracking Moving Objects Using Combined Video And Audio Information in Video Conferencing and Other Applications", both assigned to the assignee of the present invention (attorney docket references PHUS000293 and PHUS000103 respectively) and incorporated by reference herein.

The stationary video camera 210 has no need for the moving parts related to known pan, tilt, or zoom operations found in a typical non-stationary video camera or a typical video camera mounting base. The pan, tilt, and zoom functions are accomplished, as necessary, by electronically mimicking these functions with the electronic pan tilt zoom system module 220. Therefore, the video conferencing system 100 of the present invention represents a high degree of simplification as compared to known video conferencing systems.

While embodiments of the present invention have been described herein for purposes of illustration, many modifications and changes will become apparent to those

skilled in the art. Accordingly, the appended claims are intended to encompass all such modifications and changes as fall within the true spirit and scope of this invention.

## CLAIMS:

1. A video conferencing system (100) comprising:  
an image pickup device (210) for generating image signals representative of an image;  
an audio pickup device (230) for generating audio signals representative of  
5 sound from an audio source; and  
a multimodal integration architecture system (270) for processing said image signals and said audio signals to determine a direction of the audio source relative to a reference point.

10 2. The video conferencing system (100) of claim 1 wherein said multimodal integration architecture system (270) further comprises:  
an audio source localization system (240);  
a computer vision person detection system (250); and  
a multimodal speaker detection system (260).

15 3. The video conferencing system (100) of claim 2, further comprising an integrated housing (110) for an integrated video conferencing system (100) incorporating the image pickup device (210), the audio pickup device (230), and the multimodal integration architecture system (270).

20 4. The video conferencing system (100) of claim 3, wherein the integrated housing (110) is sized for being portable.

25 5. The video conferencing system (100) of claim 2, further comprising an electronic pan tilt zoom system (220) for electronically manipulating the image signals to effectively provide at least one of variable pan, tilt, and zoom functions.

6. The video conferencing system (100) of claim 5, wherein the image pickup device (210) is a stationary camera (210).

7. The video conferencing system (100) of claim 5, wherein the multimodal integrated architecture system (270) provides control signals to the electronic pan tilt zoom system (220).

5 8. The video conferencing system (100) of claim 7, wherein the audio source moves relative to the reference point, the audio source localization system (240) detects the movement of the audio source, and, in response to the movement, the audio source localization system (240) causes a change in the field of view of the image pickup device  
10 (210).

9. The video conferencing system (100) of claim 5, wherein the audio pickup device (230) is comprised of an array of two microphones (231, 232).

15 10. A method comprising the steps of:  
generating, at an image pickup device (210), image signals representative of an image;  
generating, at an audio pickup device (230), audio signals representative of sound from an audio source;  
20 processing the image signals and the audio signals to determine a direction of the audio source relative to a reference point;  
manipulating the image signals to produce refined image signals; and  
outputting said refined image signals.

25 11. The method of claim 10 further comprising the steps of:  
applying said audio signals to an audio source localization system (240);  
applying said image signals to a computer vision person detection system (250);  
processing said audio signals and said image signals with a multimodal  
30 speaker detection system (260);  
generating control signals based on the determined direction of the audio source;

applying the control signals to an electronic pan tilt zoom system (220) to mimic the effect of at least one function of a movable camera, said function selected from the group consisting panning, tilting, and zooming said movable camera; and providing an output from said electronic pan tilt zoom system (220).

5

12. The method of claim 10, further comprising electronically varying a field of view of the image pickup device (210) in response to the control signals.

1/1

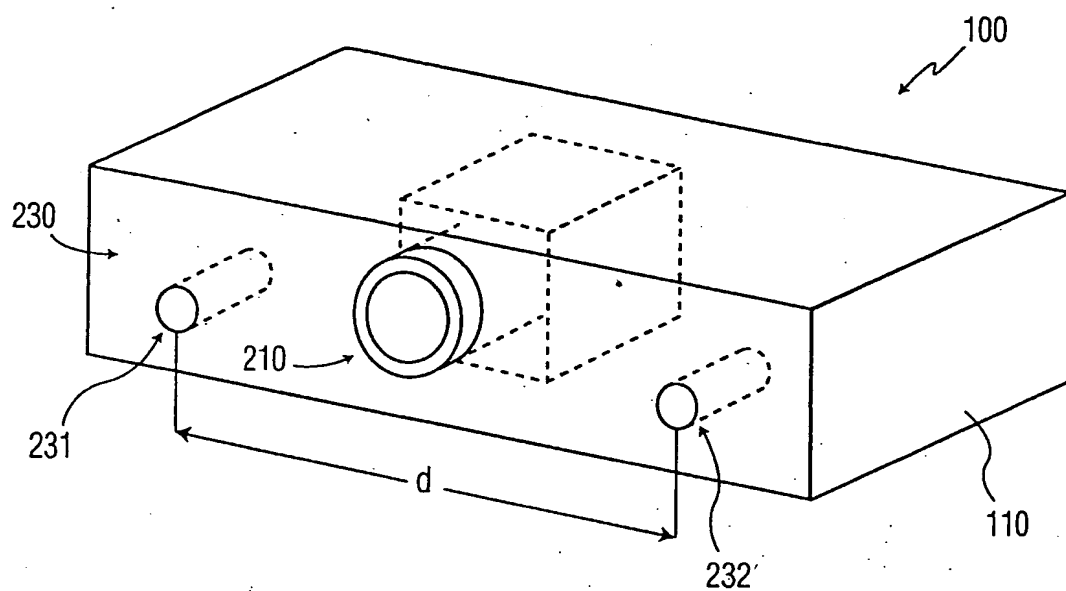


FIG. 1

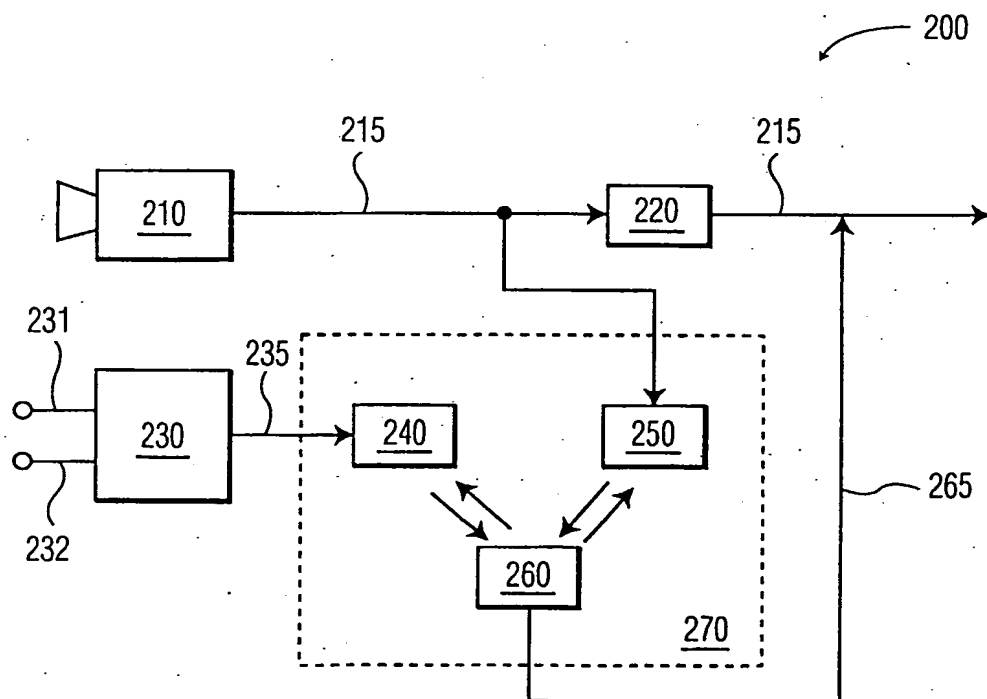


FIG. 2